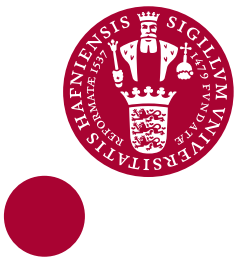


Content-Aware Video Editing in the Temporal Domain

Slot, Kristine; Truelsen, René; Sporring, Jon

Published in:
Scandinavian Conference on Image Analysis (SCIA '09)

DOI:
10.1007/978-3-642-02230-2

Publication date:
2009

Document version
Publisher's PDF, also known as Version of record

Content-Aware Video Editing in the Temporal Domain

Kristine Slot, René Truelsen, and Jon Sporring

Dept. of Computer Science, Copenhagen University,
Universitetsparken 1, DK-2100 Copenhagen, Denmark
kristine@diku.dk,rtr@rtr.dk,sporring@diku.dk

Abstract. An extension of 2D Seam Carving [Avidan and Shamir, 2007] is presented, which allows for automatic resizing the duration of a video without interfering with the velocities of the objects in the scene. We identify a set of pixels across different video frames to be either removed or duplicated in a seamless manner by analyzing 3D space-time sheets in the videos. Results are presented on several challenging video sequences.

Key words: Seam carving, video editing, temporal reduction

1 Seam Carving

Video recording is increasingly becoming a part of our every day use. Such videos are often recorded with an abundance of sparse video data, which allows for temporal reduction, i.e. reducing the duration of the video, while still keeping the important information. This article will focus on a video editing algorithm, which permits unsupervised or partly unsupervised editing in the time dimension. The algorithm shall be able to reduce, without altering object velocities and motion consistency (no temporal distortion). To do this we are not interested in cutting out entire frames, but instead in removing spatial information across different frames. An example of our results is shown in Figure 1.

Seam Carving was introduced in [Avidan and Shamir, 2007], where an algorithm for resizing images without scaling the objects in the scene is introduced. The basic idea is to constantly remove the least important pixels in the scene, while leaving the important areas untouched. In this article we give a novel extension to the temporal domain, discuss related problems and perform evaluation of the method on several challenging sequences. Part of the work presented in this article has earlier appeared as a masters thesis [Slot and Truelsen, 2008].

Content aware editing of video sequences has been treated by several authors in the literature typically by using steps involving: Extract information from the video, and determine which parts of the video can be edited. We will now discuss related work from the literature. An simple approach is frame-by-frame removal: An algorithm for temporal editing by making an automated object-based extraction of key frames was developed in [Kim and Hwang, 2000], where a key frame is a subset of still images which best represent the content of the video. The

(a)

(b) (c)

Fig. 1. A sequence of driving cars where 59% of the frames may be removed seamlessly. Frames from the original (`http://rtr.dk/thesis/videos/diku_biler_orig.avi`) is shown in (a), a frame from the shortened movie in (b) (`http://rtr.dk/thesis/videos/diku_biler_mpi_91removed.avi`), and a frame where the middle car is removed in (c) (`http://rtr.dk/thesis/videos/xvid_diku_biler_remove_center_car.avi`).

key frames were determined by analyzing the motion of edges across frames. In [Uchihashi and Foote, 1999] was presented a method for video synopsis by extracting key frames from a video sequence. The key frames were extracted by clustering the video frames according to similarity of features such as color-histograms and transform-coefficients. Analyzing a sequence as a spatio-temporal volume was first introduced in [Adelson and Bergen, 1985]. The advantage of viewing the motion using this new perspective is clear: Instead of approaching it as a sequence of singular problems, which includes complex problems such as finding feature correspondence, object motion can instead be considered as an edge in the temporal dimension. A method for achieving automatic video synopsis from a long video sequence, was published by [Rav-Acha et al., 2007], where a short video synopsis of a video is produced by calculating the activity of each pixel in the sequence as the difference between the pixel value at some time frame, t, and the average pixel value over the entire video sequence. If the activity varies more than a given threshold it is labeled as an active, otherwise as an inactive pixel at that time. Their algorithm may change the order of events, or even break long events into smaller parts showed at the same time. In [Wang et al., 2005] was an article presented on video editing in the 3D-gradient domain. In their method, a user specifies a spatial area from the source video together with an area in the target video, and their algorithm seeks optimal spatial seam between the two areas as that with the least visible transition between them. In [Bennett and McMillan, 2003] an approach with potential for differ-

ent editing options was presented. Their approach includes video stabilization, video mosaicking or object removal. Their idea differs from previous models, as they adjust the image layers in the spatio-temporal box according to some fixed points. The strength of this concept is to ease the object tracking, by manually tracking the object at key frames. In [Velho and Marín, 2007] was presented a Seam Carving algorithm [Avidan and Shamir, 2007] similar to ours. They reduced the videos by finding a surface in a three-dimensional energy map and by remove this surface from the video, thus reducing the duration of the video. They simplified the problem of finding the shortest-path surface by converting the three dimensional problem to a problem in two dimensions. They did this by taking the mean values along the reduced dimension. Their method is fast, but cannot well handle crossing objects. Several algorithms exists that uses minimum cut: An algorithm for stitching two images together using an optimal cut to determine where the stitch should occur is introduced in [Kvatra et al., 2003]. Their algorithm is only based on colors. An algorithm for resizing the spatial information is presented in [Rubenstein et al., 2008]. where a graph-cut algorithm is used to find an optimal solution, which is slow, since a large amount of data has to be maintained. In [Chen and Sen, 2008] is presented an algorithm for editing the temporal domain using graph-cut, but they do not discuss letting the cut uphold the basic rules determined in [Avidan and Shamir, 2007], which means that their results seems to have stretched the objects in the video.

2 Carving the Temporal Dimension

We present a method for reducing video sequences by iteratively removing spatio-temporal sheets of one voxel depth in time. This process is called carving, the sheets are called seams, and our method is an extension of the 2D Seam Carving method [Avidan and Shamir, 2007]. Our method may be extended to simultaneously carving both spatial and temporal information, however we will only consider temporal carving.

We detect sheets whose integral minimizes an energy function, and the energy function is based on the change of the sequence in the time direction:

$$E_1(r,c,t) = \left|\left| \frac{I(r,c,t+1) - I(r,c,t)}{1} \right|\right|, \tag{1}$$

$$E_2(r,c,t) = \left|\left| \frac{I(r,c,t+1) - I(r,c,t-1)}{2} \right|\right|, \tag{2}$$

$$E_{g(\sigma)}(r,c,t) = \left|\left| \left(I \star \frac{dg_\sigma}{dt} \right)(r,c,t) \right|\right|. \tag{3}$$

The three energy functions differ by their noise sensitivity, where E_1 is the most and $E_{g(\sigma)}$ is the least for moderate values of σ. A consequence of this is also that the information about motion is spread spatially proportionally to the objects speeds, where E_1 spreads the least and $E_{g(\sigma)}$ the most for moderate values of σ. This is shown in Figure 2

(a) (b) (c)

Fig. 2. Examples of output from (a) E_1, (b) E_2, and (c) $E_{g(0.7)}$. The response is noted to increase spatially from left to right.

To reduce the video's length we wish to identify a seam which is equivalent to selecting one and only one pixel from each spatial position. Hence, given an energy map $E \in \mathbb{R}^3 \to \mathbb{R}$ we wish to find a seam $S \in \mathbb{R}^2 \to \mathbb{R}$, whos value is the time of each pixel to be removed. We assume that the sequence has (R, C, T) voxels. An example of a seam is given in Figure 3.

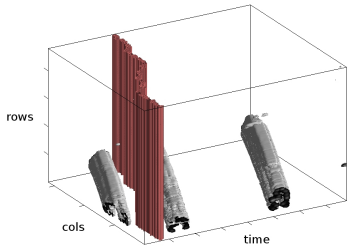


Fig. 3. An example of a seam found by choosing one and only one pixel along time for each spatial position

To ensure temporal connectivity in the resulting sequence, we enforce regularity of the seam by applying the following constraints:

$$|S(r,c) - S(r-1,c)| \leq 1 \wedge |S(r,c) - S(r,c-1)| \leq 1 \wedge |S(r,c) - S(r-1,c-1)| \leq 1. \tag{4}$$

We consider an 8-connected neighborhood in the spatial domain, and to optimize the seam position we consider the total energy,

$$\mathbb{E}_p = \min_S \left(\sum_{r=1}^{R} \sum_{c=1}^{C} E(r,c,S(r,c))^p \right)^{\frac{1}{p}}. \tag{5}$$

A sheet intersecting an event can give visible artifacts in the resulting video, wherefore we use $p \to \infty$, and terminate the minimization, when $\mathbb{E}_\infty$ exceeds a break limit b. Using these constraints, we find the optimal sheet as:

1. Reduce the spatio-temporal volume E to two dimensions.

2. Find a 2D seam on the two dimensional representation of E.
3. Extend the 2D seam to a 3D seam.

Firstly, we reduce the spatio-temporal volume E to a representation in two dimensions by projection onto either the $\mathcal{RT}$ or the $\mathcal{CT}$ plane. To distinguish between rows with high values and rows containing noise when choosing a seam, we make an improvement to [Velho and Marín, 2007], by using the variance

$$M_{\mathcal{CT}}(c,t) = \frac{1}{R-1}\sum_{r=1}^{R}(E(r,c,t) - \mu(c,t))^2. \tag{6}$$

and likewise for $M_{\mathcal{RT}}(r,t)$. We have found that the variance is a useful balance between the noise properties of our camera and detection of outliers in the time derivative.

Secondly, we find a 2D seam $p_{.\mathcal{T}}$ on $M_{.\mathcal{T}}$ using the method described by [Avidan and Shamir, 2007], and we may now determine the seam of least energy of the two, $p_{\mathcal{CT}}$ and $p_{\mathcal{RT}}$.

Thirdly, we convert the best 2D seam p into a 3D seam, while still upholding the constraints of the seam. In [Velho and Marín, 2007] the 2D seam is copied, implying that each row or column in the 3D seam S is set to p. However, we find that this results in unnecessary restrictions on the seam, and does not achieve the full potential of the constraints for a 3D seam, since it is areas of high energy may not be avoided. Alternatively, we suggest to create a 3D seam S from a 2D seam p by what we call *Shifting*. Assuming that we are working with the case of having found $p_{\mathcal{CT}}$ is of least energy, then instead of copying p for every row in S, we allow for shifting perpendicular to r as follows:

1. Set the first row in S to p in order to start the iterative process. We call this row $r = 1$.
2. For each row r from $r = 2$ to $r = R$ we determine which values are legal for the row r while still upholding the constraints to row $r - 1$ and to the neighbor elements in the row r.
3. We choose the legal possibility which gives the minimum energy in E and insert in the 3D seam S in the r'th row.

The method of Shifting is somewhat inspired from the *sum-of-pairs Multiple Sequence Alignment (MSA)* [Gupta et al., 1995], but our problem is more complicated, since the constraints must be upheld to achieve a legal seam.

3 Carving Real Sequences

By locating seams in a video, it is possible to both reduce and extend the duration of the video by either removing or copying the seams. The consequence of removing one or more seams from a video is that the events are moved close together in time as illustrated in Figure 4.

In Figure 1 we see a simple example of a video containing three moving cars, reduced until the cars appeared to be driving in convoy. Manual frame removal

(a) (b)

Fig. 4. Seams have been removed between two cars, making them appear to have driven with shorter distance. (a) Part of the an original frame, and (b) The same frame after having removed 30 seams.

may produce a reduction too, but this will be restricted to the outer scale of the image, since once a car appears in the scene, then frames cannot be removed without making part of or the complete cars increase in speed. For more complex videos such as illustrated in Figure 5, there does not appear to be any good seam to the untrained eye, since there are always movements. Nevertheless it is still

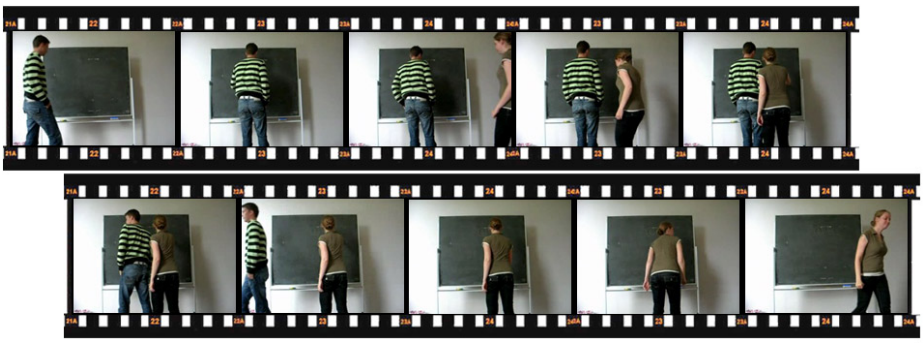


Fig. 5. Two people working at a blackboard (`http://rtr.dk/thesis/videos/events_overlap_orig_456f.avi`), which our algorithm can reduce by 33% without visual artifacts (`http://rtr.dk/thesis/videos/events_overlap_306f.avi`).

possible to remove 33% of the video without visible artifacts, since the algorithm can find a seam even if only a small part of the characters are standing still.

Many consumer cameras automatically sets brightness during filming, which for the method described so far introduces global energy boosts, luckily, this may be detected and corrected by preprocessing: If the brightness alters through the video, an editing will create some undesired edges as illustrated in Figure 6(a)(a), because the pixels in the current frame are created from different frames in the original video. By assuming that the brightness change appears somewhat evenly throughout the entire video, we can observe a small spatial neighborhood φ of the video, where no motion is occurring, and find an adjustment factor $\Delta(t)$ for each frame t in the video. If $\varphi(t)$ is the color in the neighborhood in the frame

t, then we can adjust the brightness to be as in the first frame by finding

$$\Delta(t) \;=\; \varphi(1) - \varphi(t),$$

and then subtract $\Delta(t)$ from the entire frame t. This corrects brightness problem as seen in Figure 6(b)(b).

(a) The brightness edge is visible between the two cars to the right.

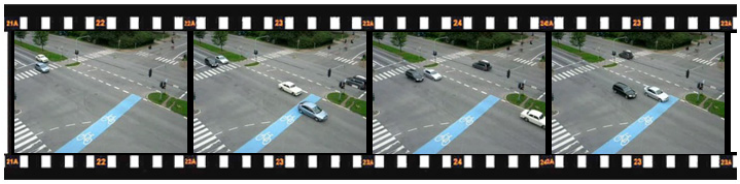
(b) The brightness edge is corrected by our brightness correction algorithm.

Fig. 6. An illustration of how the brightness edge can inflict a temporal reduction, and how it can be reduced or maybe even eliminated by our brightness correction algorithm.

For sequences with many co-occuring events, it becomes seemingly more difficult to find good cuts through the video. E.g. when objects appear that move in opposing directions, then no seams may exist that does no violate our constraints. E.g. in Figure 7(a), we observe an example of a road with cars moving in opposite directions, whos energy map consists of perpendicular moving objects as seen in Figure 8(a). In this energy map it is impossible to locate a connected 3D seam without cutting into any of the moving objects, and the consequence can be seen in Figure 7(b), where the car moving left has been stretched. For this particular traffic scene, we may perform *Spatial Splitting*, where the sequence is split into two spatio temporal volumes, which is possible if no event crosses between the two volume boxes. A natural split in the video from Figure 7(a) will be between the two lanes. We now have two energy maps as seen in Figure 8, where we notice that the events are disjunctive, and thus we are able to easily find legal seams. By stitching the video parts together after editing an equal number of seams, we get a video as seen in Figure 7(c), where we both notice that the top car is no longer stretched, and at the same time that to move the cars moving right drive closer.

4 Conclusion

By locating seams in a video, it is possible to both reduce and extend the duration of the video by either removing or copying the seams. The visual outcome, when removing seams, is that objects seems to have been moved closer together.



(a)

(b)

(c)

Fig. 7. Six selected frames from the original video (a) (`http://rtr.dk/thesis/videos/diku_crossing_243f.avi`), a seam carved video with a stretched car (b), and a seam carved video with spatial split applied (c) (`http://rtr.dk/thesis/videos/diku_crossing_142f.avi`).

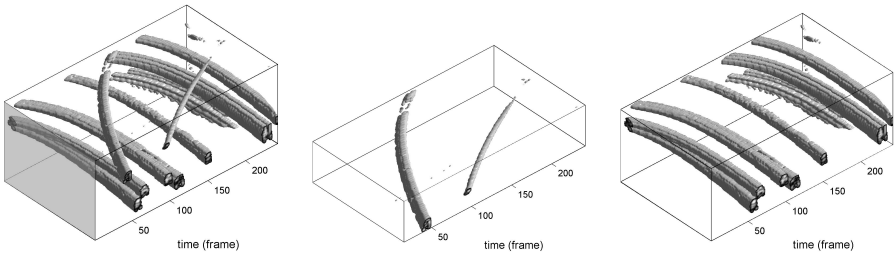
Likewise, if we copy the seams, then we will experience that the events are moved further apart in time.

We have developed a fast seam detection heuristics called *Shifting*, which presents a novel solution for minimizing energy in three dimensions. The method does not guarantee a local nor global minimum, but the tests have shown that the method is still able to deliver a stable and strongly reduced solution.

Our algorithm has worked on gray scale videos, but may easily be extended to color by (1)–(3). Our implementation is available in Matlab, and as such only a proof of concept not useful for handling larger videos, and even with a translation into a more memory efficient language, a method using a sliding time window is most likely needed for analysing large video sequences, or the introduction of some degree of user control for artistic editing.

References

[Adelson and Bergen, 1985] Adelson, E. H. and Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *J. of the Optical Society of America A*,



(a) The energy map of the video in Figure 7(a).

(b) The top part of the split box.

(c) The bottom part of the split box.

Fig. 8. When performing a split of a video we can create energy maps with no perpendicular events, thus allowing much better seams to be detected.

2(2):284–299.

[Avidan and Shamir, 2007] Avidan, S. and Shamir, A. (2007). Seam carving for content-aware image resizing. *ACM Trans. Graph.*, 26(3).

[Bennett and McMillan, 2003] Bennett, E. P. and McMillan, L. (2003). Proscenium: a framework for spatio-temporal video editing. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 177–184, New York, NY, USA. ACM.

[Chen and Sen, 2008] Chen, B. and Sen, P. (2008). Video carving. *Short Papers Proceedings of Eurographics*.

[Cormen et al., 2001] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001). *Introduction to Algorithms*. MIT Press, 2 edition.

[Gupta et al., 1995] Gupta, S. K., Kececioglu, J. D., and Schffer, A. A. (1995). *Combinatorial Pattern Matching*, chapter Making the shortest-paths approach to sum-of-pairs multiple sequence alignment more space efficient in practice, pages 128–143. Springer Berlin / Heidelberg.

[Kim and Hwang, 2000] Kim, C. and Hwang, J. (2000). An integrated sceme for object-based video abstraction. *ACM Multimedia*, pages 303–311.

[Kreyszig, 2005] Kreyszig, E. (2005). *Advanced Engineering Mathematics, 9th Edition*. John Wiley.

[Kvatra et al., 2003] Kvatra, V., Schödl, A., Essa, I., Turk, G., and Bobick, A. (2003). Graphcut textures: Image and video synthesis using graph cuts. *ACM Transactions on Graphics*, 22(3):277–286.

[Rav-Acha et al., 2007] Rav-Acha, A., Pritch, Y., and Peleg, S. (2007). Video synopsis and indexing. *Proceedings of the IEEE*.

[Rubenstein et al., 2008] Rubenstein, M., Shamir, A., and Avidan, S. (2008). Improved seam carving for video editing. *ACM Transactions on Graphics (SIGGRAPH)*, 27(3):to appear.

[Slot and Truelsen, 2008] Slot, K. and Truelsen, R. (2008). Content-aware video editing in the temporal domain. Master's thesis, Dept. of Computer Science, Copenhagen University. www.rtr.dk/thesis.

[Uchihashi and Foote, 1999] Uchihashi, S. and Foote, J. (1999). Summarizing video using a shot importance measure and a frame-packing algorithm. In *the International Conference on Acoustics, Speech, and Signal Processing (Phoenix, AZ)*, volume 6,

pages 3041–3044, FX Palo Alto Laboratory, 3400 Hillview Avenue, Palo Alto, CA 94304.

[Velho and Marín, 2007] Velho, L. and Marín, R. D. C. (2007). Seam carving implementation: Part 2, carving in the timeline. `http://w3.impa.br/~rdcastan/SeamWeb/Seam\%20Carving\%20Part\%202.pdf`.

[Wang et al., 2005] Wang, H., Xu, N., Raskar, R., and Ahuja, N. (2005). Videoshop: A new framework for spatio-temporal video editing in gradient domain. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, page 1201, Washington, DC, USA. IEEE Computer Society.